Sommaire

1 State of the art

2 Formulation F

3 Scaling up through clustering

4 Project

General information

Group

- Alone or in pairs

Programming language

- Free

  Julia code provided

Schedule

- 31/03 : due date

Clustering

Requested work

1. Apply F to the 3 provided datasets
   Code provided
2. Apply F_U, F_S^e and F_S^h
   Code provided (naive clustering method)
3. Apply F, F_U, F_S^e and F_S^h to two other datasets
   You can find some here : https ://archive.ics.uci.edu/ml/datasets.php
4. Address at least one of the following open questions :
   1. Design and test other clustering methods
   2. Theoretical clusering results similar to Property 2
   3. Identification and use of valid inequalities
   4. Use a lazy callback to reduce the size of the formulation
   5. Design and test other separations shifting algorithms
   6. Any idea which enables to improve the computation time or the accurracy

Remark

In the provided code, the split function are centered a posteriori to be as far as possible from the data

In order to limit overfitting

Open question 1 : Other clustering methods

Naive algorithm provided

```
Data :
  $\{(x_i, y_i)\}_{i\in\mathcal{I}}$ : dataset
  $\gamma \in [0, 1]$ : percentage of data reduction
Result :
  $\mathcal{C}$ : data partition
$\mathcal{C} \leftarrow \{C_i = \{i\}\}_{i\in\mathcal{I}}$ tant que $\underline{|\mathcal{C}| \geq \gamma|\mathcal{I}|}$ faire
  Merge the clusters $C$ and $C'$ of the same class which minimize $\min_{i\in C} \min_{i'\in C'} ||x_i - x_{i'}||_2$
```

Requested work

Design other data clustering algorithms which :

- does not completely ignore the class of the data ; and
  Example : do not simply apply the *k*-means on all data without taking into account their class
- does not treat each class independently
  Example : do not simply apply one *k*-means for the data of each class

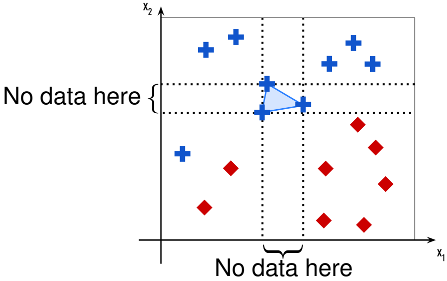
Compare your performances to the ones of the provided methods

Algorithms which enable to obtain clusters with data from different class (to better handle outliers) could be interesting.

Open question 2 : Clustering theoretical results

Requested work

Find one or several clustering hypothesis similar to H_1 which enable to obtain optimal or near-optimal solutions. It is allowed to add additional conditions on the data (ideally not too-constraining ones)



Open question 3 : Valid inequalities

Context

The linear relaxation of MILP for optimal classification trees is bad

High gap at the root

Requested work

Identify and add valid inequalities to the formulation

- Statically (when building the MILP) or
- Dynamically (during the resolution)

  In a callback to cut fractional points

Evaluate the efficiency

Remark

Inequalities that you did not find by yourself and which do not provide an improvement will not be considered sufficient

Open question 4 : Use a lazy callback to reduce the size of the formulation

Context

The size of the formulation grows quickly with $|\mathcal{I}|$

Requested work

Initially remove inequalities from the formulation and generate them when necessary

In a callback to cut integer solutions

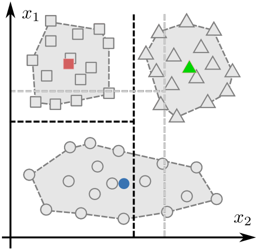
Evaluate the efficiency

Open question 5 : Improve the shifting algorithm

Requested work

Design a new approach to shift the split functions in the iterative algorithm
Evaluate if your method enables to

- reduce the global computation time ;
- reduce the number of iterations ; or
- obtain better solutions with F_S^h.



Open question 6 : Any other idea

Requested work

Find and test any ideas which can improve the performances